Required Programs and Data Files

Intro to Next-Generation Sequencing Workshop

Botany 2010

August 1, 2010; Providence, RI

Please download and install the following programs on your laptop prior to the workshop. Note that the minimum requirement for the Genomics Workbench is 2 GB RAM. If you have < 2 GB RAM, you should still try the program. If the program does not function or is taking too long to open the data files, you should plan to work with another attendee during the hands on portion of the workshop.

A. Software to Install

1. CLC bio Genomics Workbench

Follow the download and activation process instructions for a free two-week trial of the CLC Genomics Workbench at http://www.clcbio.com/index.php?id=87.

2. Tablet

Follow the download instructions for the Tablet next-generation sequence assembly viewer at http://bioinf.scri.ac.uk/tablet/download.shtml.

3. 7-zip

Windows users will need to download the 7-Zip file archiver (free) available at http://www.7-zip.org/ if they do not already have a utility capable of unpacking Gzip (.gz) files. This capability is already included in the Archive Utility on Macs.

B. Download sequence files in fastq format from NCBI Sequence Read Archive.

1. Go to GenBank Sequence Read Archive - http://www.ncbi.nlm.nih.gov/sra.
2. In the search field, type in the following: Pinus thunbergii
3. Click on record SRX014012 (this should be the second record).
4. On the right side of the screen, click on “Download data for this experiment SRX014012”.
5. Download the file **SRR030730.fastq.gz** to your computer
6. Go back to the Sequence Read Archive using the back button/arrow on your browser, and search for *Pinus lambertiana*.
7. Click on record SRX014023 (this should be the first record).

8. On the right side of the screen, click on “Download data for this experiment SRX014023”.

9. Download the file **SRR027096.fastq.gz** to your computer.

10. The downloaded files now need to be unzipped. If you are using a Mac, simply double-click on each zipped file and it should unzip. If you are using a PC/Windows, you will probably have to download a utility to unzip the files. We use and recommend the program 7-zip, available free for download at: http://www.7-zip.org/download.html.

C. Download the annotated reference sequence from GenBank.

1. From GenBank, search CoreNucleotide for *Pinus thunbergii* chloroplast, complete genome.

2. Scroll down and click on the record for the *Pinus thunbergii* chloroplast genome (NC_001631.1 GI:7524593).

3. In the upper right of your screen, click on the highlighted ‘Send’, and select Complete Record: File: Format GenBank. Then ‘Create File’ and download to your computer. The file should save as sequences.gb.

D. Uploading the read and reference files into CLC Genomics Workbench.

1. Open CLC Genomics Workbench.

2. In the navigation area (upper left), right-click (windows) or ctrl+click (Mac) on the folder CLC_Data and create a new folder called Reads and Reference.

3. Repeat the previous step twice to make two more folders called De Novo Assembly and Reference-Guided Assembly.

4. From the dropdown menu, select File: Import High-Throughput Sequencing Data: Illumina.